Bayesian inference for infectious disease modelling

Summer School Halle 2022

Fabienne Krauer (LSHTM)

cmmid centre *for* mathematical modelling *of* infectious diseases

LONDON SCHOOL *of* HYGIENE &TROPICAL MEDICINE

Recommended Readings

Literature

Bayesian Data Analysis (Third edition).
Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin.
Online here: https://users.aalto.fi/~ave/BDA3.pdf

Statistical Rethinking (Second edition). Richard McElreath.

Bayesian workflow. ArXiv 2020. Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, Martin Modrák. https://arxiv.org/abs/2011.01808

Online resources

- https://avehtari.github.io/BDA_course_Aalto/
- http://florianhartig.github.io/LearningBayes/
- Stan user guide https://mc-stan.org/docs/stan-users-guide/index.html
- MCMC samplers: https://m-clark.github.io/docs/ld_mcmc/
- https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html
- Guided example for model fitting in Stan/rstan: https://mc-stan.org/users/documentation/case-studies/boarding_school_case_study.html
- https://chi-feng.github.io/mcmc-demo/app.html?algorithm=NaiveNUTS&target=banana
- https://m-clark.github.io/docs/ld_mcmc/

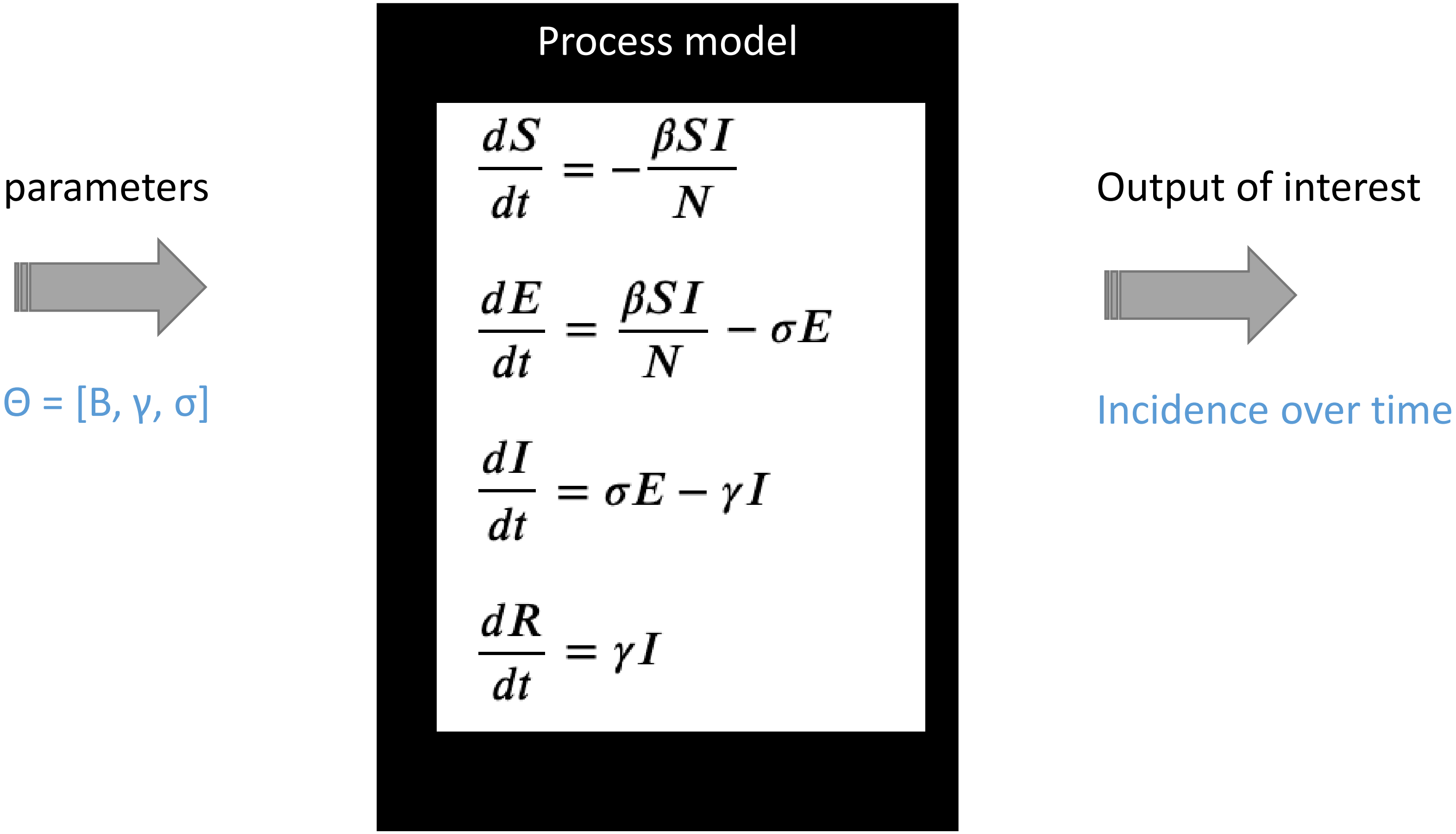
Table of content

- Inference and Likelihood
- Bayesian theorem and inference
- Sampling algorithms
- The Bayesian workflow
- R packages for fitting ODE models

“model”

“A simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions”.

Oxford English Dictionary



parameters

Θ = [B, γ, σ]

Process model

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \sigma E$$

$$\frac{dI}{dt} = \sigma E - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

Output of interest

Incidence over time

“model fitting”

1. What does it mean to “fit a model”?

To estimate the magnitude of the parameters in the process model that best describe the observed data

2. WHY?

To learn something about the disease system

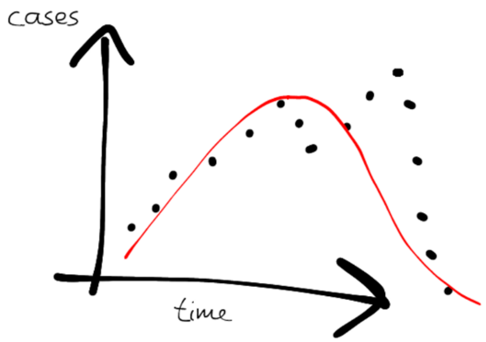
To use the results for forecasting/prediction in other situations

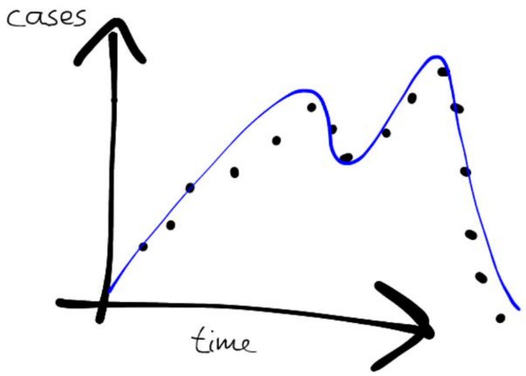
3. HOW?

By matching the outcome of interest of the process model as close as possible to the data

→ typically, the outcome of interest is an incidence time series (“trajectory matching”)

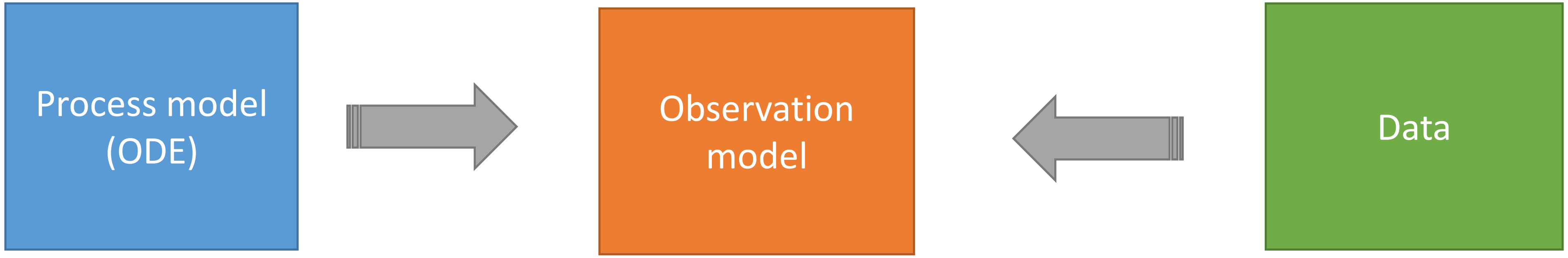
→ Alternative, we can use a summary estimate (e.g. cumulative number of cases per year, see ABC)

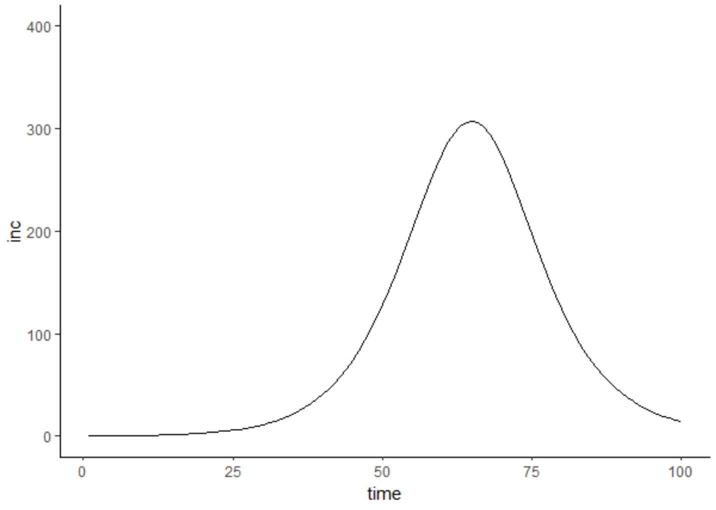




Graphs by Sebastian Funk, Model fitting short course LSHTM

Linking the model and data: likelihood-based inference

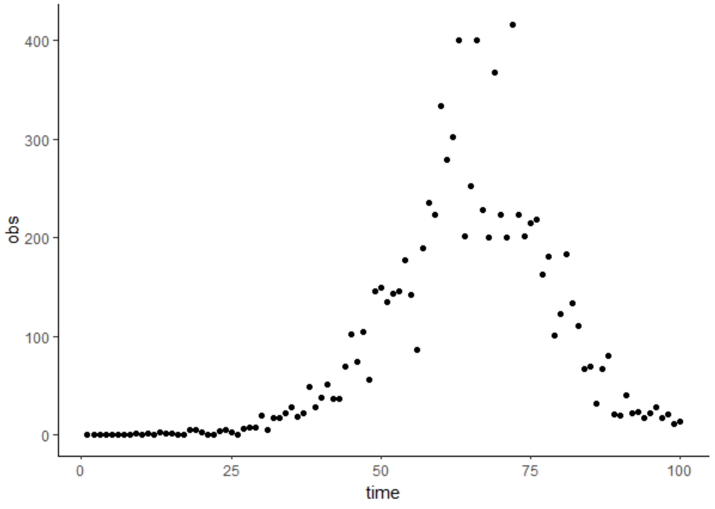




Measures the uncertainty in the outcome of interest from the process model

=

likelihood-based inference



“Inference” and “likelihood”

“**Statistical inference** is the process of using data analysis to infer properties of an underlying distribution of probability”

Oxford Dictionary of Statistics

“The **likelihood function** (often simply called the likelihood) describes the joint probability of the observed data as a function of the parameters of the chosen statistical model”

Casella, Statistical Inference, 2002

$$\Pr(Data \mid parameters) = \Pr(y|\theta)$$

The likelihood is NOT a probability density function in the strict sense (i.e. does not integrate to 1)

Likelihood in practice

1) Choose a likelihood function that describes the measurement error best:

Data type	**Data example**	**Likelihood function**	**Property**	**R function**

Count data	Incidence	Poisson	Mean = Variance	dpois(x, lambda)
				rpois(n, lambda)

		Negative Binomial	Mean < Variance (=overdispersion)	dnbinom(x, size, mu)
				rnbinom(1, size, mu)

Proportions	seroprevalence	Binomial		dbinom(x, size, prob)
				rbinom(1, size, prob)

Likelihood in practice

2) Apply the chosen likelihood function to a data point and the corresponding model point.

Example: Poisson distribution

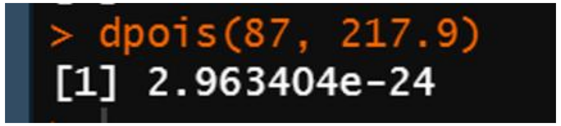
At t = 56
Observation (“x”) = 87 cases
Model prediction (“λ”) = 217.9 cases

“by hand”:

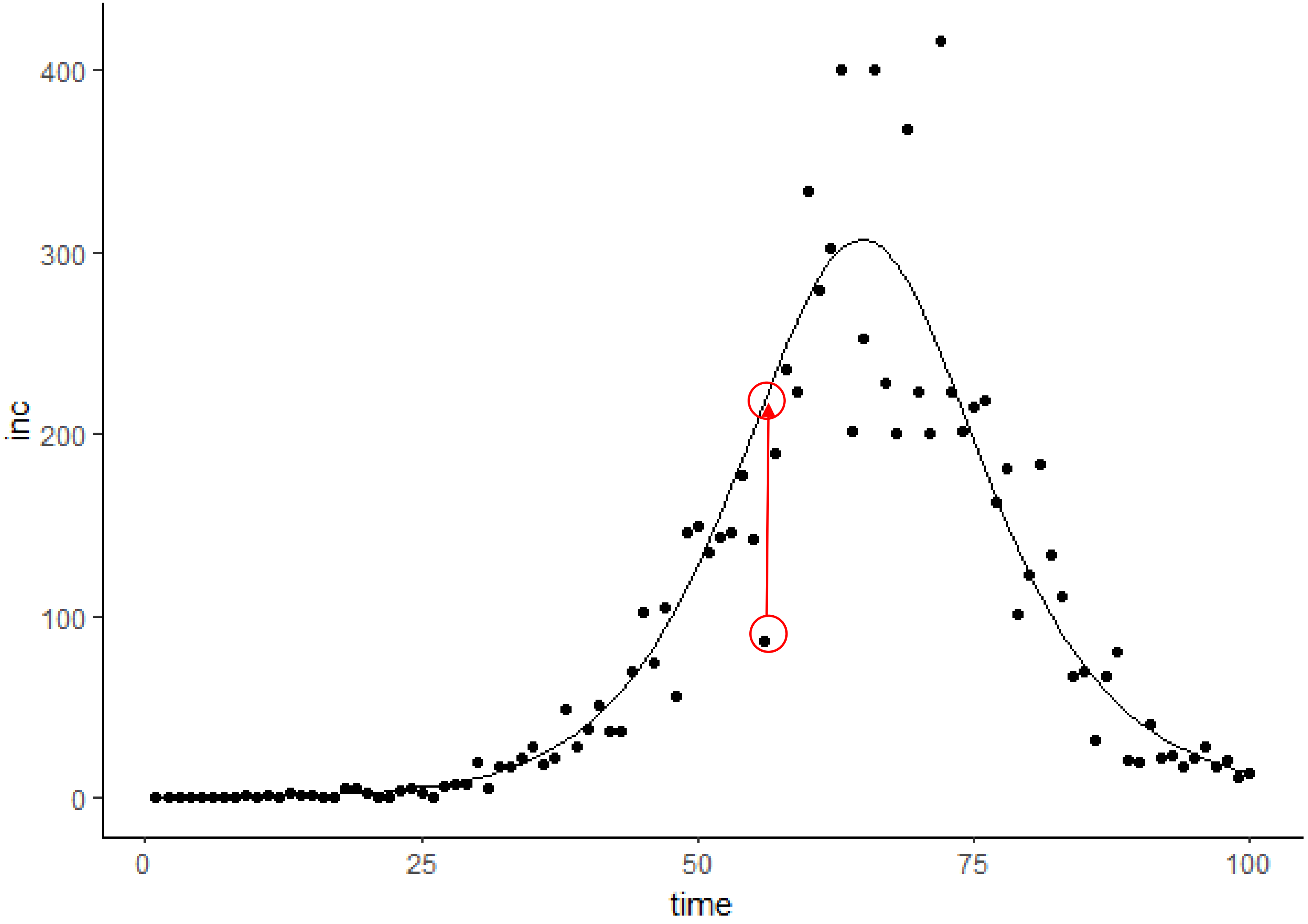
$$\Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

P(data point| theta) = 2.963404e-24

In R:



```
> dpois(87, 217.9)
[1] 2.963404e-24
```



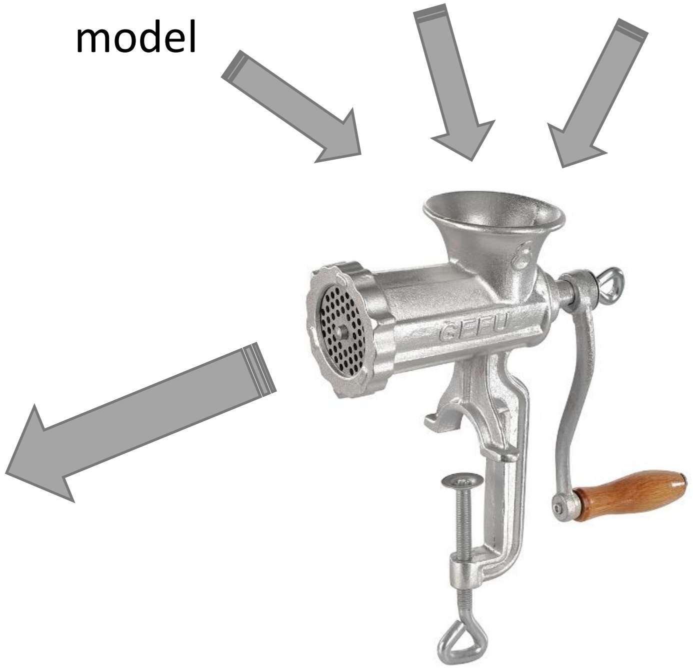
Likelihood in practice

3) Log transform the resulting probability (In R: dpois(87, 217.9, log=TRUE))

4) Repeat this for every data point / model point and sum the resulting probabilities (“joint probability” remember?), which yields the log likelihood:

$$\log(\Pr(data|\theta)) = \sum_i \log(p(data\ point\ i\ |\theta))$$

$$\Pr(data|\theta) = \prod_i p(data\ point\ i\ |\theta)$$



Observation model

Data

Process model (ODE)

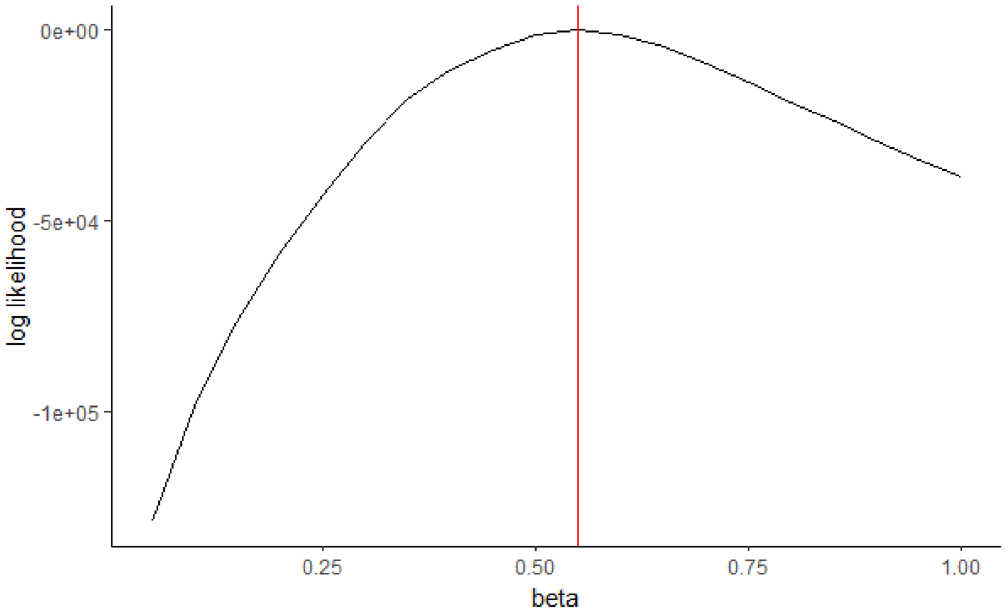
Log likelihood
-3518.18

Linking the likelihood and inference

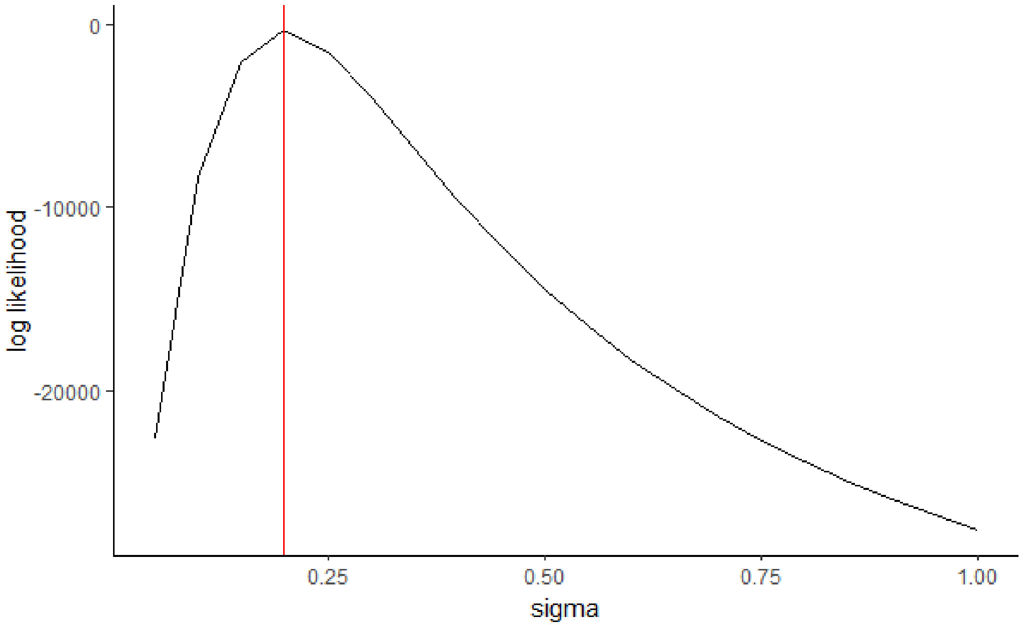
Example: ODE model with β=0.6 and σ=0.2

What are the log likelihood values if we vary these parameters? → likelihood profile

Fix σ, vary β from 0 to 1

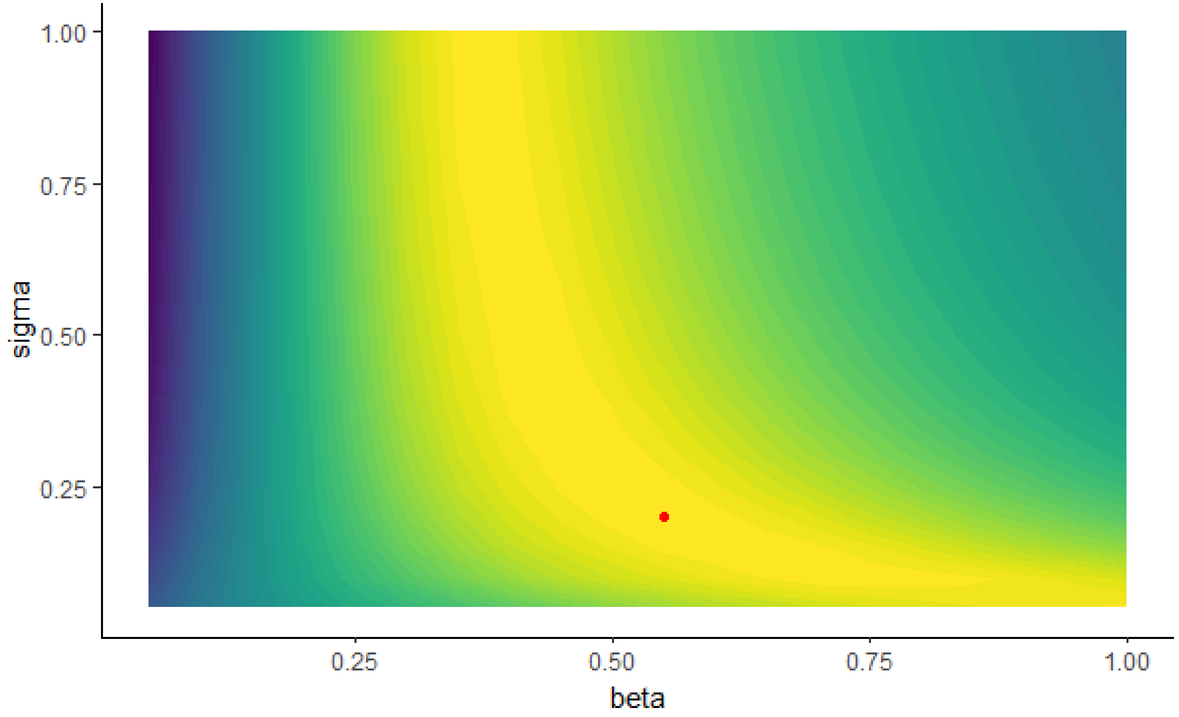


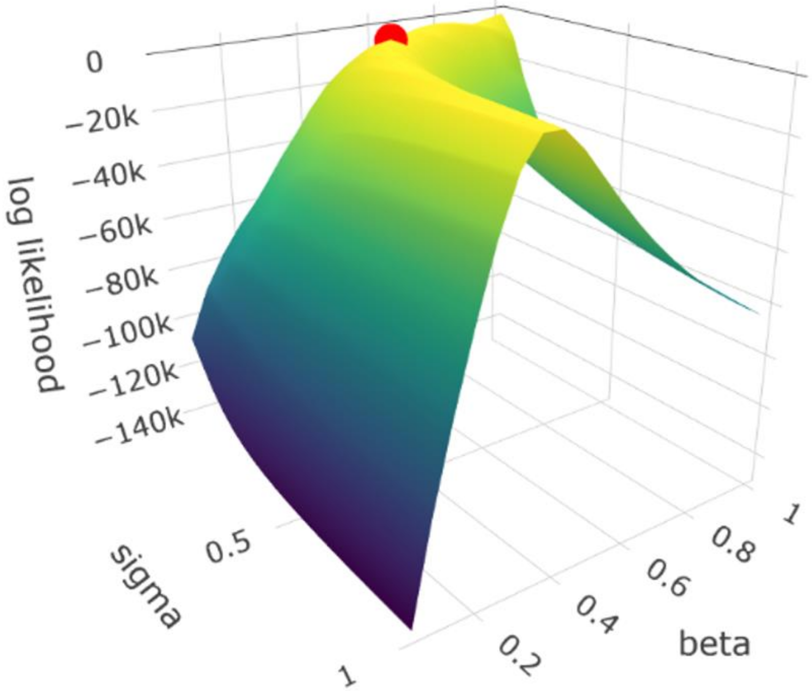
Fix β, vary σ from 0 to 1



Linking the likelihood and inference

Inference explores the log likelihood landscape (surface) while varying all the parameters of interest (the fitted parameters)





Likelihood derived inference methods

	Null hypothesis testing	**Maximum likelihood estimation (MLE)**	**Bayesian inference**
Uncertainty	data	data	parameters
output	Rejection of H0 or not	MLE point estimator (set of theta that maximizes the likelihood)	Posterior distribution
Confidence	P-value	Confidence intervals	Credible intervals

Frequentist statistics

Adapted from Florian Hartig’s slides, Bayesian Summer School 2017

Bayesian theorem

“Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations. “

Gelman, Bayesian Data Analysis.

Thomas Bayes (1763). An Essay Towards Solving a Problem in the Doctrine of Chances

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

$$\Pr(disease|test\ positive) = \frac{\Pr(test\ postitive|disease)\Pr(disease)}{\Pr(test\ positive)}$$

$$\mathrm{PPV} = \frac{sensitivity\ * prevalence}{(sensitivity\ * prevalence) + (1 - specificity) * (1 - prevalence)}$$

Bayesian theorem & Bayesian inference

$$\Pr(\vartheta|y) = \frac{\Pr(y|\theta)\,\Pr(\vartheta)}{\Pr(y)}$$

Likelihood

= process model + observation model

→ likelihood of the observed data (y) under a model and parameter set (ϑ)

Prior

Suggestions for distributions for the parameter set ϑ BEFORE having seen the data y ("a-priori belief")

Posterior

= Joint posterior probability distribution

→ Expresses uncertainty about the parameter set ϑ after taking both the data and the prior into account

Marginal likelihood (normalizing constant)

→ Indicates what the data theoretically can look like BEFORE the actual data has been observed (using only the prior distribution and the log likelihood function).

→ Calculated by integrating/summing the joint probability distribution over the parameter space

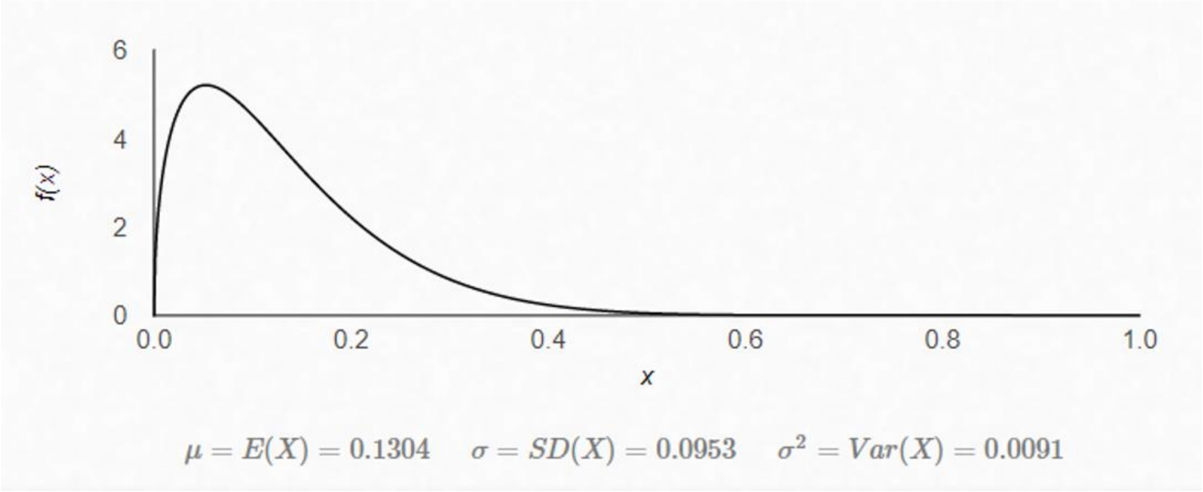
→ Purpose: to normalize the posterior to a proper probability distribution (so that it integrates to 1)

“Prior”

- Encodes your “prior belief” of what the parameter distribution looks like
- Can and should influence the posterior
- The more data, the less influence has the prior
- Choose appropriate distribution:
- Beta() for proportions (0,1),
- Gamma() or LogNormal() for positive parameters
- Uniform() for uninformative priors without prior knowledge except the bounds

- Example:

- $\gamma \sim Beta(1.5, 10.0)$



Why Bayesian?

1) Bayesian is more intuitive:

a) Bayesian credible interval: 95% probability, that the true estimate lies in this range
b) Frequentist confidence interval: in 95% of cases, the true estimate lies within the specified range

2) Bayesian provides more information about the distribution of the parameters (rather than a point estimate and some CI

3) Bayesian acknowledges that there is uncertainty in parameters

4) Bayesian let’s you update your model with new information (the prior)

The three steps of Bayesian inference

1) Setting up a full probability model:
    a) Define the process model
    b) Define an observation model
    c) Define the priors of the parameters of interest

2) Conditioning on observed data (aka fitting)

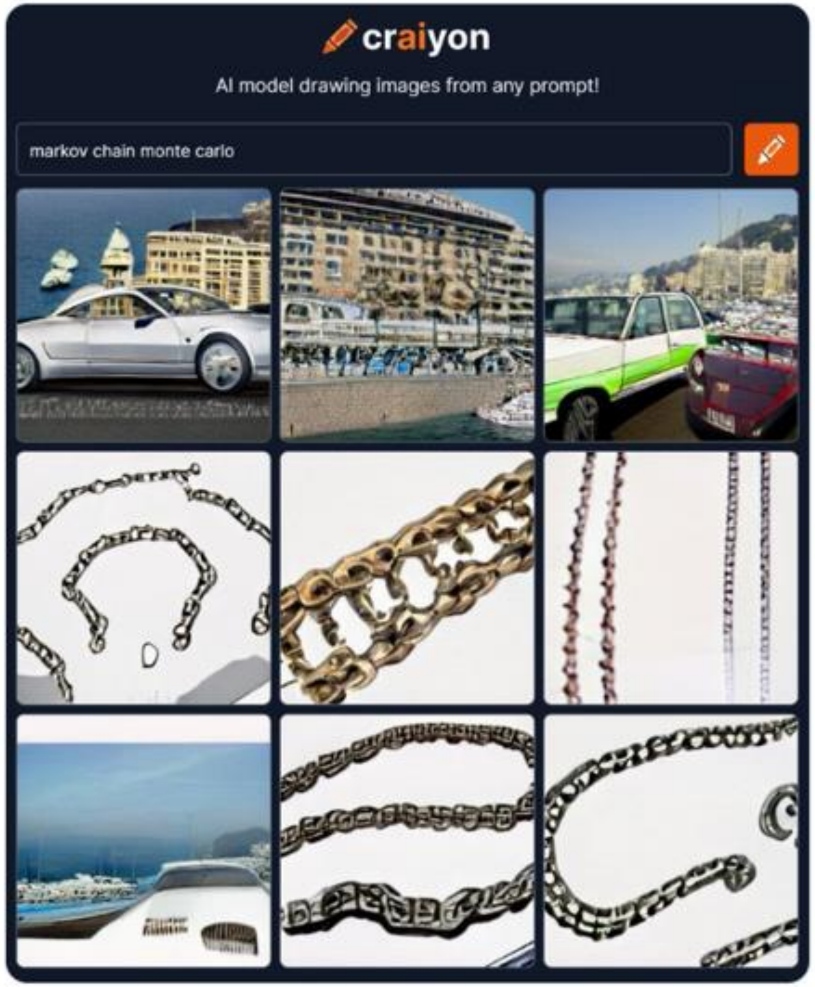
1) Evaluating the fit
    a) Evaluate the computation (convergence)
    b) Evaluate the fit to the data

Bayesian inference “by hand”: a practical example

Analytical solution vs. numerical solution

- Analytically calculating the marginal likelihood is difficult or even impossible for a large parameter set (requires integrals for continuous parameters)
- Calculating the profile log-likelihood for every possible value of the parameter is computationally very expensive
- We need a numerical (computational) approximation.
- We can use sampling strategies to sample (“draw”) from the posterior (rather than calculating it analytically)

Sampling algorithms: Markov Chain Monte Carlo MCMC



"Monte Carlo"

Class of methods for repeated random sampling from a distribution

"Markov Chain"

A sequence of states, where each state depends only on the previous state

The "MCMC" algorithm generates a sequence of samples ("markov chain")

With an increasing number of iterations (draws), the distribution of the proposed parameters (theta) more and more resembles the desired target (posterior) distribution → sampling from the posterior

"Metropolis-Hastings MH"

A particular algorithm of using MCMC to sample from a distribution

Sampling algorithms: MH-MCMC

1. Start from an initial parameter sample θ and evaluate the target function ("current" target)
2. Draw a new parameter sample θ' from a pre-defined parameter space ("the proposal distribution") and evaluate the target function for the new θ' ("proposed target"): Multivariate Normal
3. Compare the current and proposed target function:

$$ratio = \frac{f(\vartheta\prime)}{f(\vartheta)} = \frac{posterior(\vartheta')}{posterior(\vartheta)} = \frac{\frac{\Pr\left(y\middle|\theta\right)\prime\Pr(\vartheta\prime)}{\Pr(y)}}{\frac{\Pr\left(y\middle|\theta\right)\Pr(\vartheta)}{\Pr(y)}} = \frac{\Pr\left(y\middle|\theta'\right)\Pr(\vartheta\prime)}{\Pr\left(y\middle|\theta\right)\Pr(\vartheta)}$$

4. Calculate the acceptance ratio $r = \min(1, ratio)$
5. Draw a random uniform number u(0,1)
6. If u < r → accept θ' as the new current θ, else keep the current θ

Rinse and repeat 2-6 for n iterations

Graph by Sebastian Funk, Model fitting short course LSHTM

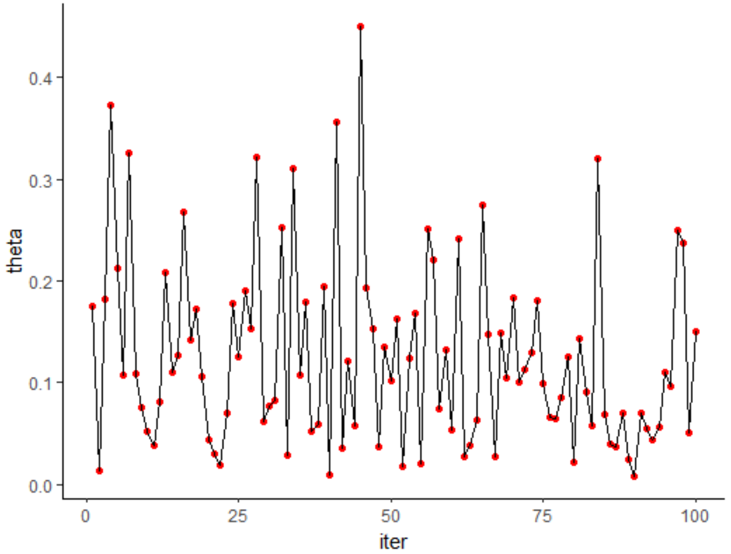
Sampling algorithms: MH-MCMC

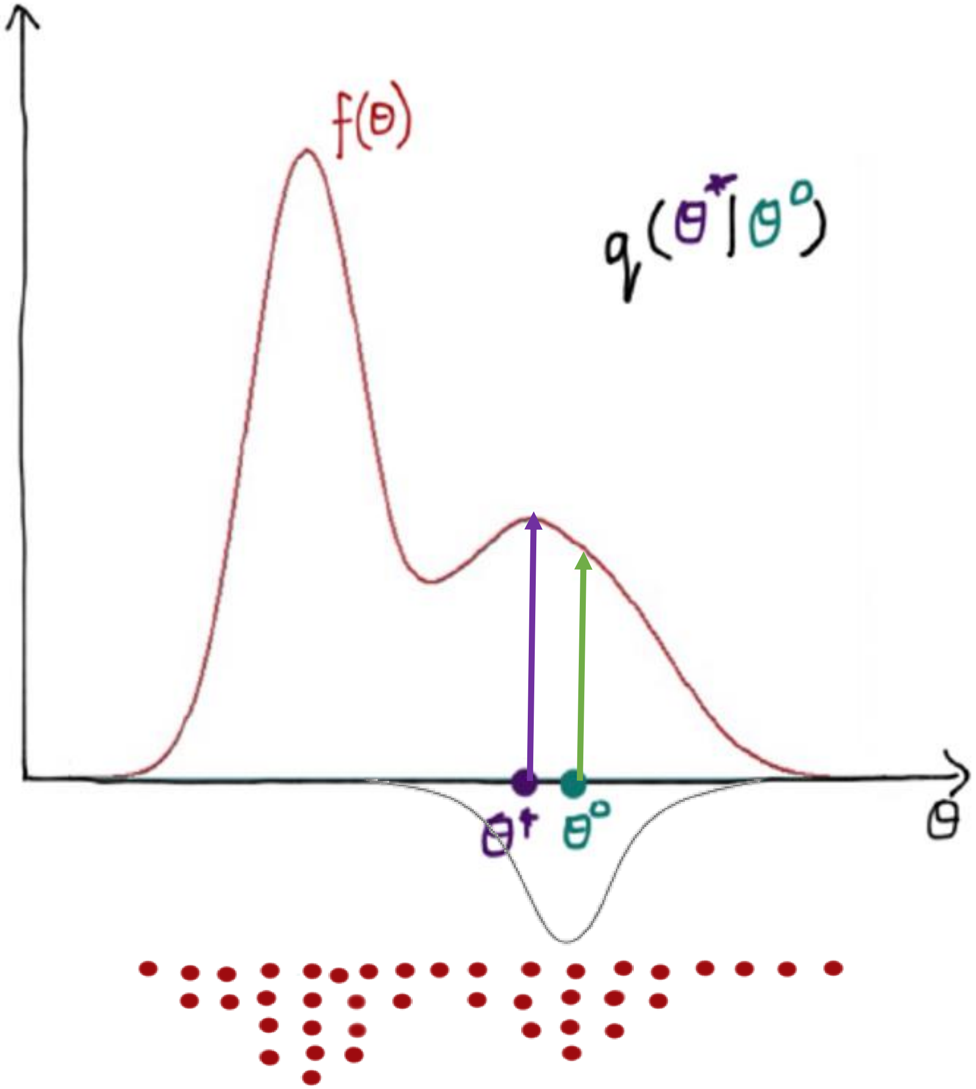
1. Draw θ and evaluate the posterior
2. Draw θ’ from a multivariate normal and evaluate the posterior
3. Ratio of current and proposed posterior

4. Acceptance ratio $r = \min(1, \frac{proposal\ posterior}{current\ posterior})$

6. Draw a random uniform number u(0,1)

7. If u < r → accept θ’ as the new current θ, else keep the current θ

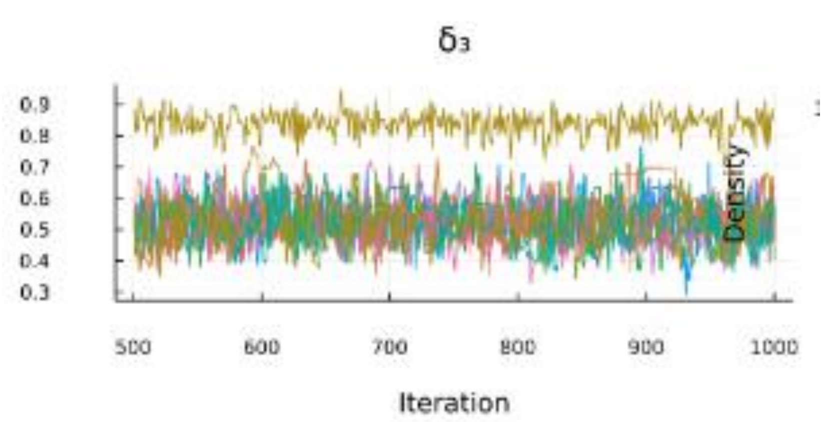


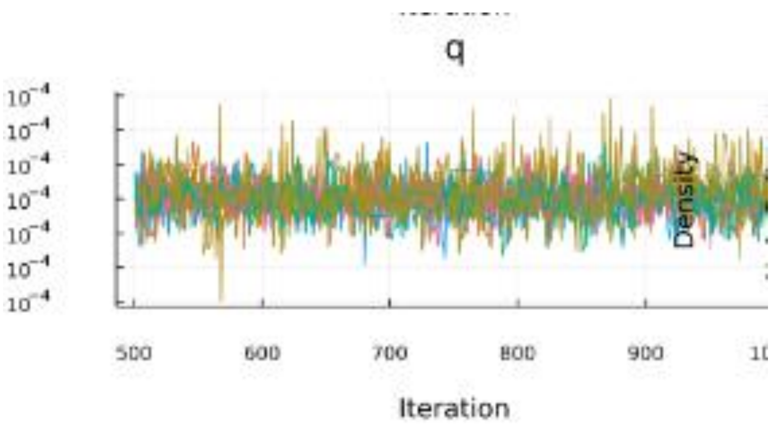


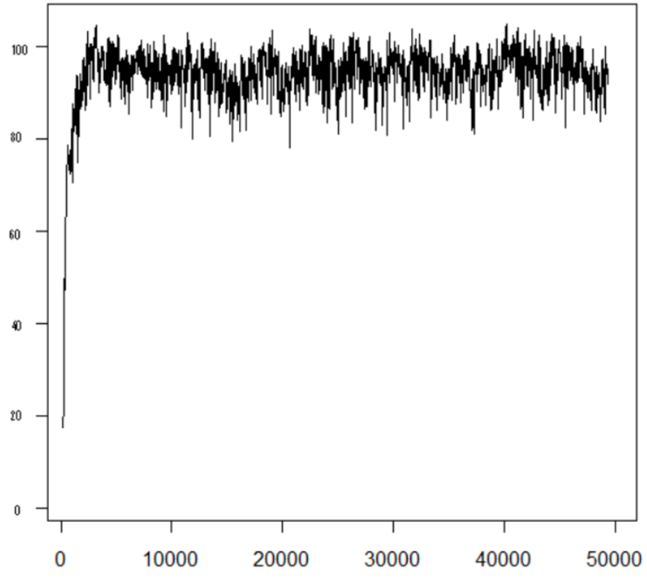
Convergence diagnostics

- Convergence = Samples have reached a stationary distribution, which consists of samples of the posterior → fuzzy caterpillar

- Need multiple chains converging to the same posterior to fully confirm convergence
- Formal assessment of convergence with Rhat / PSRF: ratio of between and within chain variance. Rhat < 1.05 indicative of convergence
- Burn-in phase must be removed
- Aim for an acceptance rate of ~23% (for random walks)



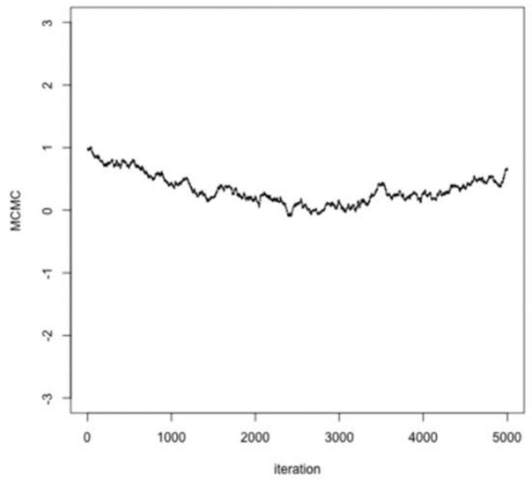




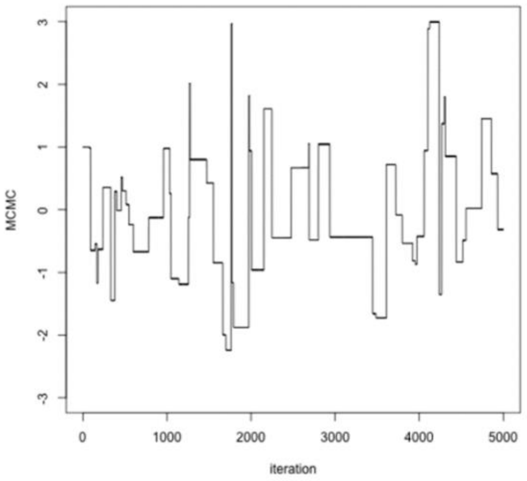
Convergence diagnostics

Good chain mixing produces a fuzzy caterpillar

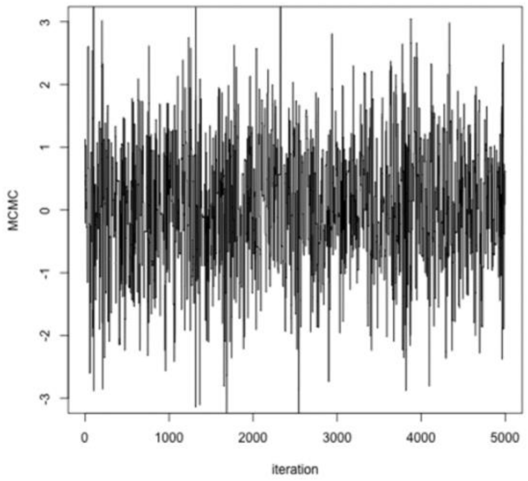
Proposal small → small step size → high acceptance rate, poor mixing



Proposal large → large step size → low acceptance rate, poor mixing



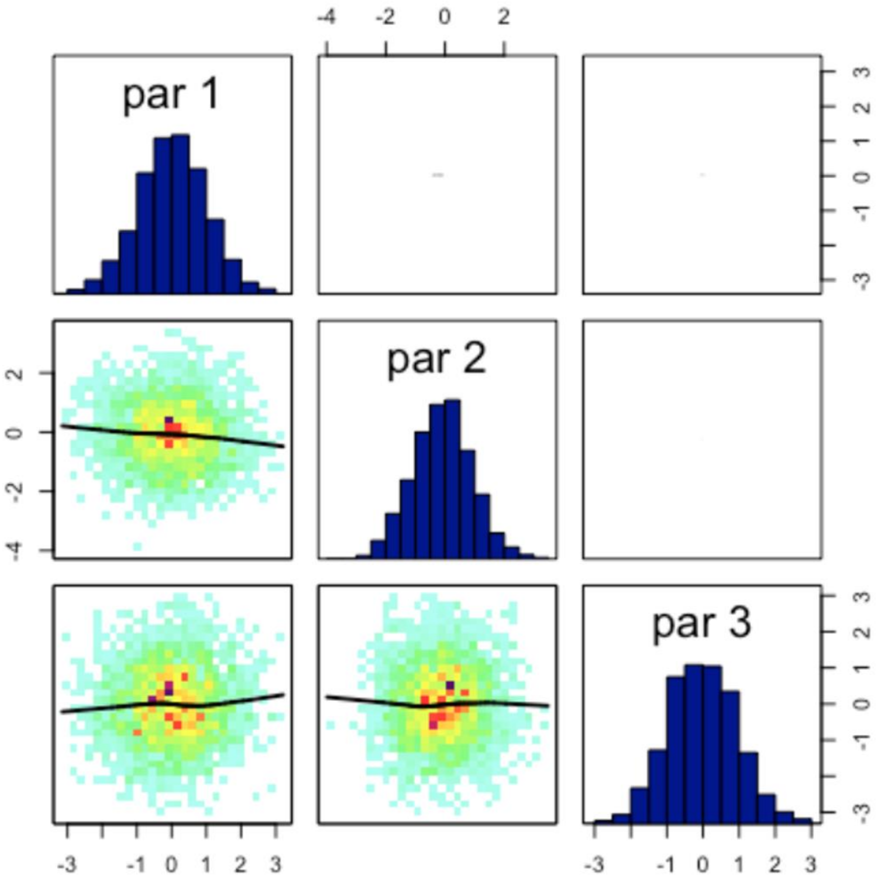
Optimal proposal → hairy catterpillar

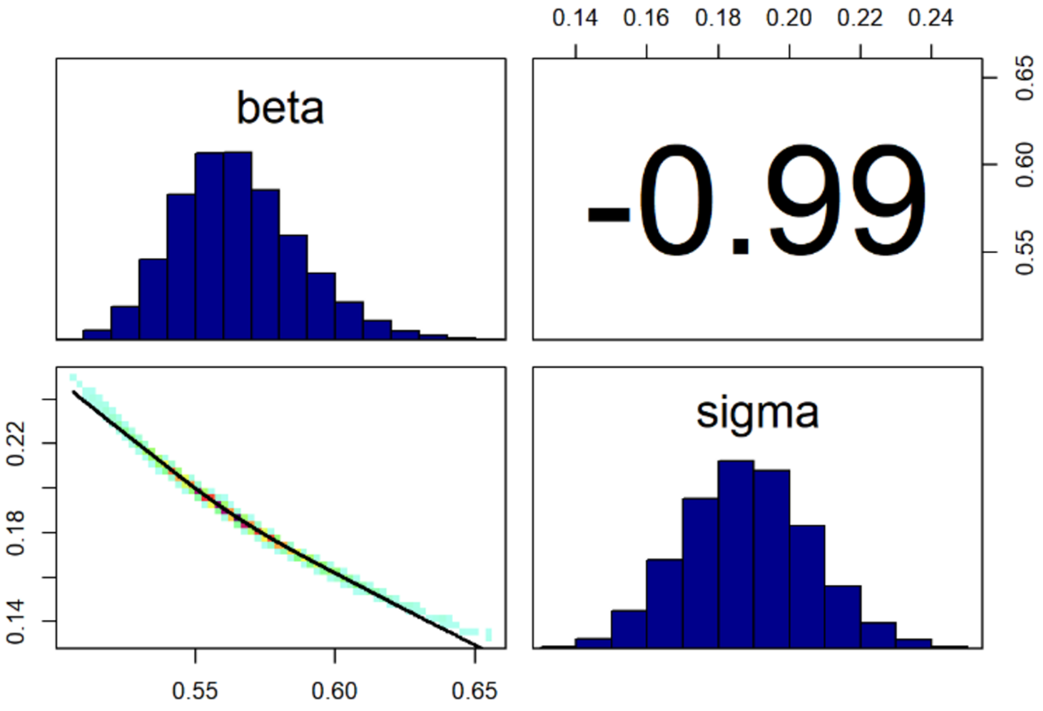


Effective sample size (ESS) indicates the number of independent samples (free from autocorrelation) → should be at least 100.

Marginal posterior statistics

Visualize parameter correlation

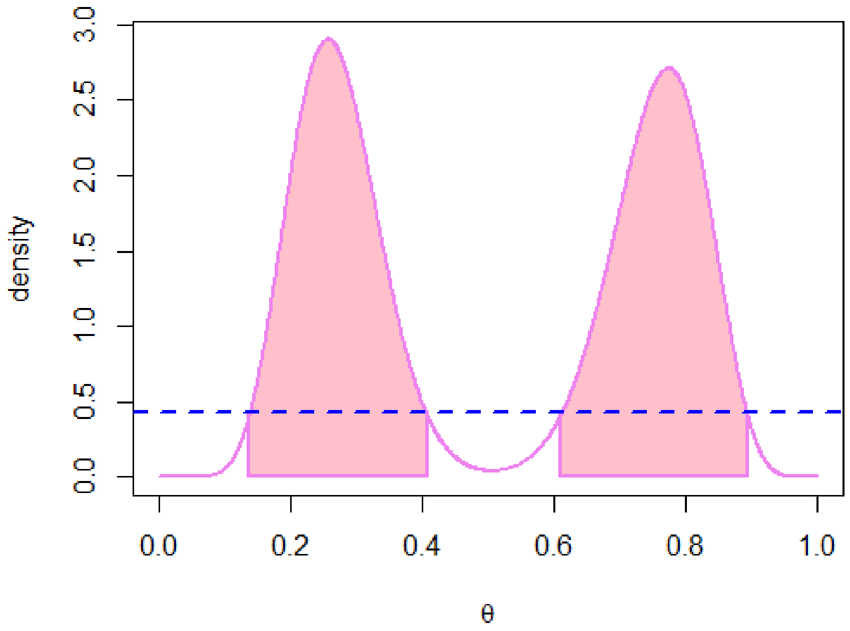


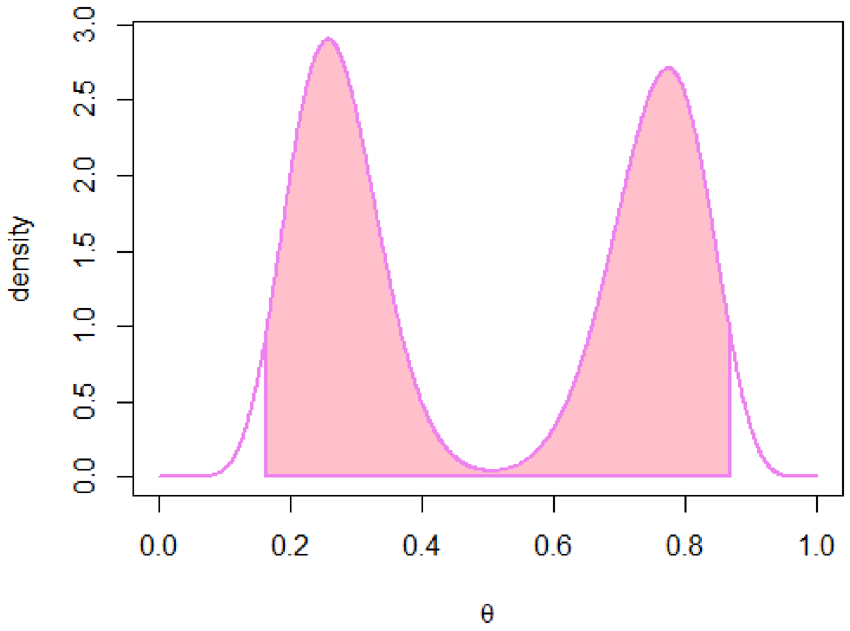


Marginal posterior statistics

Summarize the posterior distributions with 95% credible intervals (contains the true parameter with 95% probability)

- 95% highest posterior density interval (HPDI) (narrowest interval that contains 95% of the data)
- Equal tailed 95% interval (2.5th and 97.5th quantile)





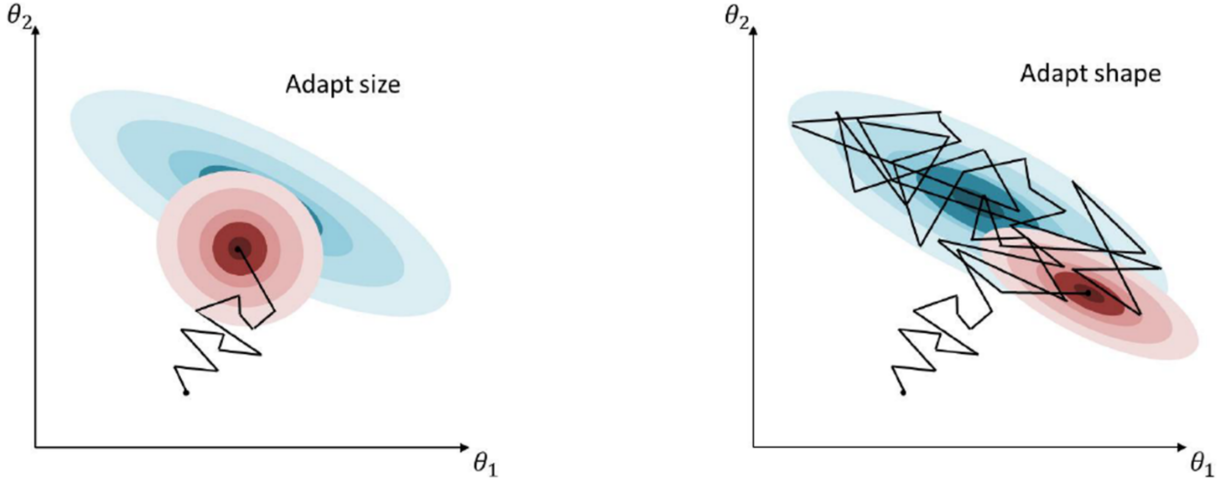
Graph by Sebastian Funk, Model fitting short course LSHTM

Adaptive MH-MCMC

- Standard MH-MCMC is very sensitive to choice of proposal (difficult to get good mixing)
- Poor performance with multiple parameters that are correlated

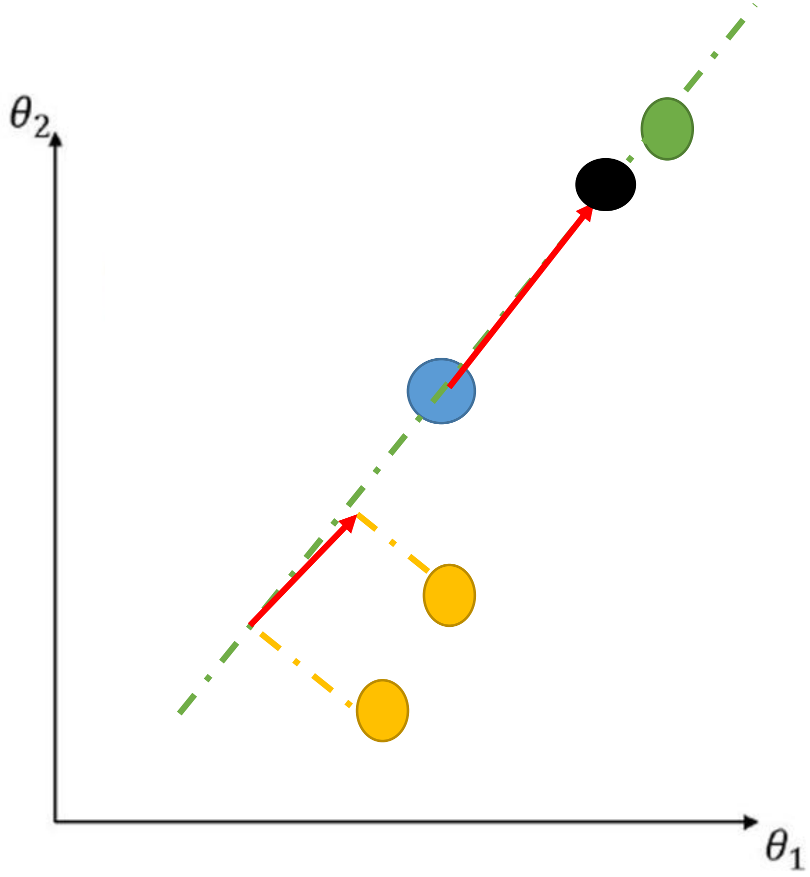
- Adaptive MH: alters the proposal during the burn-in to match the covariance matrix of the posterior. First step size (by varying the standard deviation), then shape (by varying the covariance between the parameter)

Two-stage adaptation



Sampling algorithms: DE-MCMCzs

Group of 3 chains, proposal for chain 1 is informed by the current and past state of the other two chains (no assumption of multivariate normal)



1. Choose current theta and evaluate the posterior

2. Project a vector to a randomly chosen state of another chain (current or past)

3. Select two other states randomly and project their distance onto the first projection line

4. The resulting difference vector is multiplied by some factor and added to the original point to draw the final proposal

5. The proposal is accepted with some probability

Rinse and repeat for all three chains

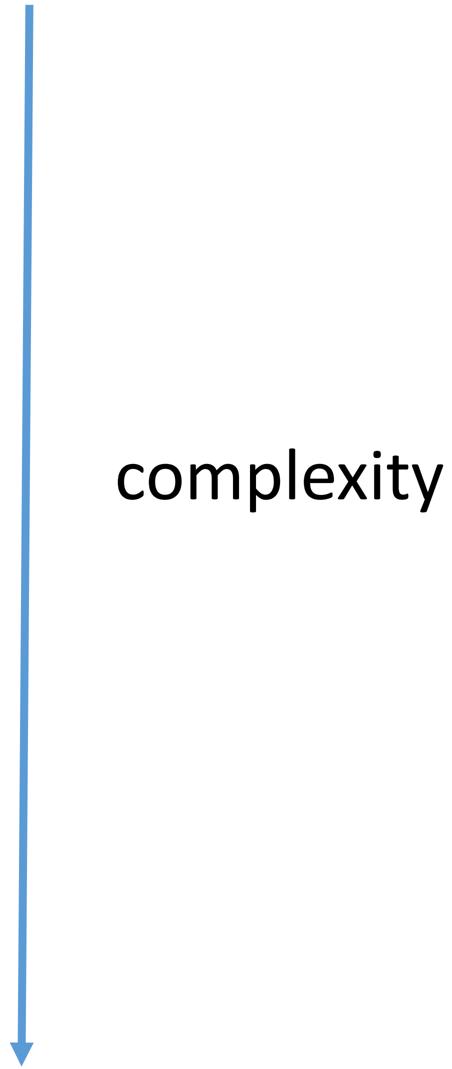
Ter Braak et al, Differential Evolution Markov Chain with snooker updater and fewer chains. Stat Comput 2008

Choice of sampling algorithms

MH-MCMC
Adaptive MH-MCMC and other variants

Genetic algorithms (DE)
Parallel tempering

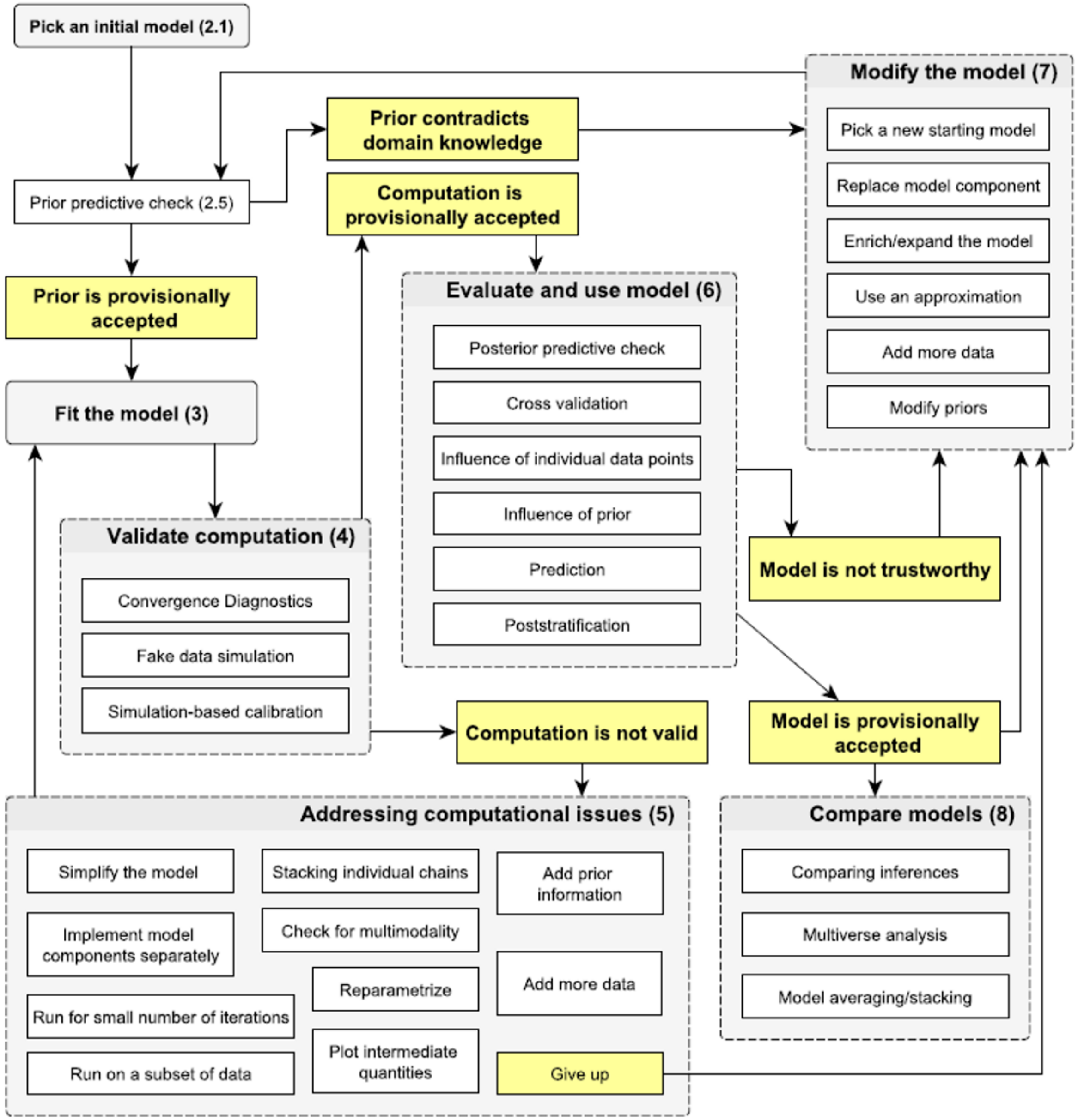
Hamiltonian Monte Carlo / NUTS



complexity

The Bayesian workflow

Gelman et al. Bayesian workflow. ArXiv 2020



Gelman et al. Bayesian workflow. ArXiv 2020

Prior predictive checks

Prior predictive checks generate data from the prior distribution to check whether the prior is appropriate:

1. Sample multiple times from the prior
2. plug the values into the process model and simulate the trajectory
3. visualize all the resulting trajectories together with the actual data

--> If the prior predictions do not overlap with actual data, you need to adjust the prior

Trouble shooting

1. Choose a sampler appropriate for your model (HMC/NUTS/DE/parallel tempering >>> MH). Don’t write your own sampler.
2. Follow the Bayesian workflow!
3. Start with the simplest model and the fewest parameters to fit. Gradually increase the complexity
4. Start by simulating data from the model and trying to fit the simulated model
5. Parameter correlation/non-identifiability makes convergence difficult:
   - Consider adding additional data (hospital data in addition to times series data) and a join likelihood
   - Adjust/adapt/tighten your priors
   - Fix one of the correlated parameters if you can
6. Complex posteriors (local modes, abrupt changes) make convergence difficult:
   - Choose initial starting values for theta by sampling from the prior OR
   - Run MLE on the model first, and use the MLE as initial values for the MCMC

R packages for Bayesian ODE inference

Package	Samplers	Advantages	Disadvantages
LaPlacesDemon	Metropolis variants Genetic variants Gibbs HMC variants T-walk, SMC	- Wealth of samplers to choose from - Fast!	- Few example code published for ODE-IDD models
BayesianTools	Metropolis variants Genetic variants SMC T-walk	- Requires to write out the log likelihood function and the prior functions (good for learning Bayes)	- MH and DE not very performant for complex models (poor mixing, no convergence). - Not very fast
Stan/Rstan	HMC/NUTS	- Very fast (C++) - Gradient-based samplers good for complex problems - Great documentation - Support from user forum	- Stan DSL (awkward way of coding the model and the components) - Steep learning curve
pomp	pMCMC	- Fast?	- Pomp DSL (awkward way of coding the model)

Acknowledgments

- Sebastian Funk, LSHTM. Material from Bayesian Model fitting for infectious disease modelling → next course summer 2023

- Florian Hartig, Uni Regensburg. Material from Bayesian Model fitting with Bayesian Tools

- Yang Liu, LSHTM. For testing the exercises.